Learning Efficient Feature Representation for Temporal Action Localization

Team: Dahua_001

Chenglu Wu*, Xuefeng Yang*, Fuzhi Duan, Yanxun Yu, Yayun Wang, Jun Yin

Zhejiang Dahua Technology Co., Ltd.
Temporal action localization (TAL) requires to precisely locate the temporal boundaries of action instances and accurately classify the action instances into specific categories.

[1]. Challenge and Workshop on Localized and Detailed Understanding of Human Actions in Videos [https://deeperaction.github.io/fineaction/]
II. Evaluation Criteria

Interpolated Average Precision (AP) is used as the metric for evaluating the results on each activity category. Then, the AP is averaged over all the activity categories (mAP). The official metric used in competition is the **average mAP**, which is defined as the mean of all mAP computed with tIoU thresholds between 0.5 and 0.95 with a step of 0.05.

In addition to evaluate proposal quality, Average Recall (AR) under multiple tIoU thresholds are calculated. AR@AN is defined as the AR under different average proposal numbers (AN), and we calculate the area under the AR vs. AN curve (**AUC**) as a metric of proposal.
III. Mainstream Algorithms

Video Classification.

SlowFast

CSN

TSN&NeXtVLAD

Video-swin-transformer

[2]. Christoph Feichtenhofer, etc. Slowfast networks for video recognition. ICCV 2019
[3]. Du Tran, etc. Video classification with channel-separated convolutional networks. ICCV 2019
[5]. Rongcheng Lin, etc. NeXtVLAD: An efficient neural network to aggregate frame-level features for large-scale video classification. ECCV-Workshops 2018
III. Mainstream Algorithms

Proposal Generation.

**BMN**
- Feature Extraction
- Base Module
- Temporal Evaluation Module
- Proposal Evaluation Module

Generate Proposals

Provide Confidences

Boundary-Matching Network

Proposal Refinement.

**TCAnet**
- Local Temporal Encoder
- Global Temporal Encoder
- LGTE
- TBR

Candidate Proposal

Start-Boundary Regression

Segment-Level Regression

End-Boundary Regression

Candidate Proposal

Internal Context

Starting Context

Segment-Level Regression Output

Frame-Level Regression Output

Fused Output

[7]. Tianwei Lin, etc. Bmn: Boundary-matching network for temporal action proposal generation. ICCV2019
[8]. Zhiwu Qing, etc. Temporal context aggregation network for temporal action proposal refinement. CVPR2021
IV. Our Approach

Dataset cleansing.

- **Class Skew Balance**
  - Original distribution VS. resample distribution.
  - Blue bar is the distribution of original number of different categories.
  - Red bar is the number of unbalanced video clips by resampling.

- **Outlier Identification & Cleansing**
  - Median proportion and the average length of instances.
  - Light coral bar is the median percentage of the total video duration for a single instance per category.
  - Purple bar is the average length of single instance of each category.
IV. Our Approach

Pipeline:

A. Feature Extraction.
B. Temporal Proposal
C. Proposal Refine: dense2sparse.
D. Detection Results Generation.

Learning Efficient Features Representation for TAL (overall)
IV. Our Approach

Feature Enhancement module:

- The improved base-feature layer.

- Encode local and global temporal relationships.

- Deform the receptive field and enhance the aggregation of context information.

The architecture of cascade proposals refinement:

- Pick samples corresponding to a specific iou for training.

- Boost the performance of the proposal prediction gradually.
V. Experiments Results

Classification Results:

<table>
<thead>
<tr>
<th>Method</th>
<th>SlowFast</th>
<th>CSN</th>
<th>TSN</th>
<th>NeXtVLAD</th>
<th>Video-swin-transformer</th>
</tr>
</thead>
<tbody>
<tr>
<td>backbone head</td>
<td>Res101+50</td>
<td>Res152</td>
<td>Swin-Base</td>
<td>Swin-Base</td>
<td>Swin3D-Base</td>
</tr>
<tr>
<td>clip_len</td>
<td>SlowFastHead</td>
<td>I3DHead</td>
<td>TSNHead</td>
<td>NextVLADHead</td>
<td>I3DHead</td>
</tr>
<tr>
<td>frame_interval</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>num_clips</td>
<td>1</td>
<td>1</td>
<td>32</td>
<td>32</td>
<td>1</td>
</tr>
</tbody>
</table>

*Training details for video classification networks.*

<table>
<thead>
<tr>
<th>Model</th>
<th>Top-1</th>
<th>Top-2</th>
<th>Top-3</th>
<th>Top-5</th>
</tr>
</thead>
<tbody>
<tr>
<td>TSN</td>
<td>81.54</td>
<td>91.46</td>
<td>94.19</td>
<td>96.87</td>
</tr>
<tr>
<td>SlowFast (nc=8)</td>
<td>84.92</td>
<td>94.11</td>
<td>96.47</td>
<td>98.63</td>
</tr>
<tr>
<td>CSN (nc=8)</td>
<td>85.70</td>
<td>94.39</td>
<td>96.96</td>
<td>98.78</td>
</tr>
<tr>
<td>Video-SwinB (nc=5)</td>
<td>87.61</td>
<td>94.90</td>
<td>97.41</td>
<td>98.82</td>
</tr>
<tr>
<td>Video-SwinB (nc=8)</td>
<td>87.90</td>
<td>95.23</td>
<td>97.84</td>
<td>99.04</td>
</tr>
<tr>
<td>NeXtVLAD</td>
<td>87.21</td>
<td>94.87</td>
<td>97.24</td>
<td>98.70</td>
</tr>
<tr>
<td>Ensemble</td>
<td><strong>90.03</strong></td>
<td><strong>96.72</strong></td>
<td><strong>98.43</strong></td>
<td><strong>99.41</strong></td>
</tr>
</tbody>
</table>

We used NeXtVLAD, CSN and Video-SwinB to ensemble the model.
V. Experiments Results

Proposal Results:

<table>
<thead>
<tr>
<th>Video feat</th>
<th>L</th>
<th>AR@1</th>
<th>AR@5</th>
<th>AR@10</th>
<th>AR@100</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>I3D</td>
<td>100</td>
<td>4.92</td>
<td>10.35</td>
<td>13.38</td>
<td>24.64</td>
<td>19.57</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>4.87</td>
<td>10.40</td>
<td>13.75</td>
<td>27.78</td>
<td>21.28</td>
</tr>
<tr>
<td>TSN-K700</td>
<td>200</td>
<td>5.15</td>
<td>11.44</td>
<td>15.10</td>
<td>29.61</td>
<td>23.07</td>
</tr>
<tr>
<td></td>
<td>250</td>
<td>5.05</td>
<td>10.90</td>
<td>14.41</td>
<td>28.64</td>
<td>22.04</td>
</tr>
<tr>
<td>Slowonly-k700</td>
<td>200</td>
<td>5.09</td>
<td>11.19</td>
<td>14.86</td>
<td>29.25</td>
<td>22.67</td>
</tr>
<tr>
<td>TSN-full-2048</td>
<td>200</td>
<td>5.42</td>
<td>12.15</td>
<td>16.08</td>
<td>31.29</td>
<td>24.53</td>
</tr>
<tr>
<td>SwinB-k600</td>
<td>200</td>
<td>5.80</td>
<td>12.98</td>
<td>16.94</td>
<td>31.73</td>
<td>25.05</td>
</tr>
<tr>
<td></td>
<td>256</td>
<td>5.82</td>
<td>12.90</td>
<td>17.09</td>
<td>32.26</td>
<td>25.32</td>
</tr>
<tr>
<td></td>
<td>325</td>
<td>6.21</td>
<td>14.23</td>
<td>18.75</td>
<td>35.00</td>
<td>27.69</td>
</tr>
<tr>
<td>SwinB-full-1024</td>
<td>256</td>
<td>6.07</td>
<td>13.97</td>
<td>18.60</td>
<td>33.96</td>
<td>27.28</td>
</tr>
<tr>
<td></td>
<td>325</td>
<td>6.21</td>
<td>14.44</td>
<td>19.32</td>
<td>36.84</td>
<td>29.25</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td>6.49</td>
<td>15.13</td>
<td>20.30</td>
<td>38.80</td>
<td>30.81</td>
</tr>
<tr>
<td></td>
<td>450</td>
<td>6.34</td>
<td>15.20</td>
<td>20.32</td>
<td>38.61</td>
<td>30.67</td>
</tr>
</tbody>
</table>

The L denotes the length of video feature.

- The AUC of the proposal at L400 is higher than L325 and L450 from the validation results of different temporal scales.
- The feature extracted by the SwinB model showed a significant improvement compared to the other features.
## V. Experiments Results

**Detection Results:**

<table>
<thead>
<tr>
<th>BMN</th>
<th>LGTE</th>
<th>GCNeXt</th>
<th>Dilate</th>
<th>TCAnet</th>
<th>NMS</th>
<th>Cascade</th>
<th>Average-mAP(%)</th>
<th>Promotion</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>17.32</td>
<td>-</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>17.61</td>
<td>+1.67%</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td>18.21</td>
<td>+3.41%</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>18.63</td>
<td>+2.31%</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>19.14</td>
<td>+2.74%</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>21.19</td>
<td>+10.71%</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td><strong>22.05</strong></td>
<td>+4.06%</td>
</tr>
</tbody>
</table>

*Influence of different modules on the performance of FineAction.*

- The BMN network trained with the standardized video feature length of 400 was the baseline.
- The detection result achieves 22.05% on the validation set and 23.35% on the test set in terms of average mAP.
VI. Conclusion

- We find that the model using the video clips for action recognition has a greater performance on proposals than the model using single frame.

- Increasing the grid size of BMN can further improve the recognition accuracy of short actions in the FineAction. However, with the increase of grid size, the total number of parameters of the model will increase exponentially, and the model convergence epoch will move backward as well.

- Considering the cumbersome nature of this method, we will improve the one-stage TAL framework to locate and identify the extremely short instances in the future.
THANKS